C

LAB 3 - CLASSIFICATION

IENG3304 – DATA MANAGEMENT & ANALYTICS

MAY 23, 2022

WRITTEN BY: ABILASH SURENDRAN
BANNER ID: B00891410

Table Of Content:

LIST OF FIGURES:

LIST OF TABLES:

Introduction:

The term "heart disease" refers to a variety of heart conditions. Heart disease can sometimes be "silent," with no indications or symptoms until a person has a heart attack, heart failure, or arrhythmia. Heart disease is exacerbated by high blood pressure, high cholesterol, and smoking. The dataset I used has 918 rows and 12 columns and contains information such as the patients' age, sex, type of chest pain, resting blood pressure, cholesterol, fasting blood sugar, resting electrocardiogram, maximum heart rate, exercise-induced angina, old peak, and heart disease, which is the output parameter. The linear discriminant analysis is utilised in this assignment to classify whether the patient has heart disease or not. Several LDA models are created and the best out of them is identified. The null classifiers and the success rates of each of the models are listed.

Body:

Heart disorders are serious problems that disrupt a person's daily life. A patient's cardiovascular disease is caused by several variables. Smoking, an unhealthy lifestyle, excessive alcohol consumption, and other factors are among the most prominent causes. Let's look at each of the parameters in dataset one by one.

Age: The patient's age range when they first develop the cardiac disease is mentioned.

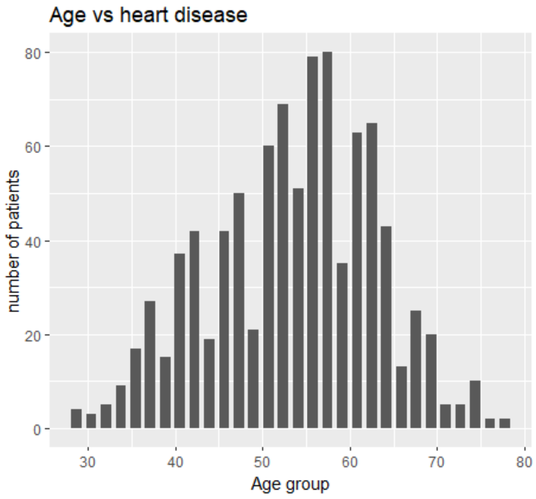

Fig 1: Histogram for Age

The chart shows that individuals between the ages of 55 and 60 are the most affected, relative to the rest of the population. Age-related changes in the heart and blood vessels can increase the risk of cardiovascular disease.

Sex: This parameter is used to analyse which gender is comparatively healthy. Here male and female gender is being classified.

Cholesterol: Cholesterol is a nutrient that contributes to heart disease. It can be fatal if not handled appropriately. Cholesterol is measured in milligrams per decilitre.

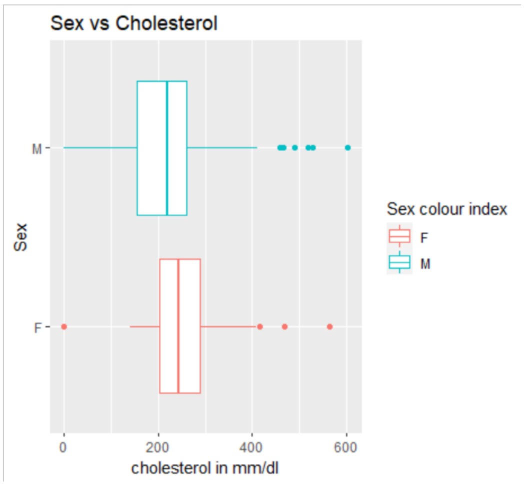

Fig 2: Sex vs cholesterol in mm/dl

The above box plot depicts the distribution of cholesterol between the two genders. The plot clearly shows that male cholesterol levels range from 0 to around 400 mm/dl. On the other hand, it varies between 180 and 410 mm/dl for females. Any value above 400 mm/dl can be noted as outliers from the plot. However, an increase in cholesterol levels increases the risk of heart disease.

<u>Chest pain type:</u> The various types of chest pains that cause heart diseases are listed below. Typical angina, atypical angina, non-anginal pain, and asymptomatic pain are the four types of pain.

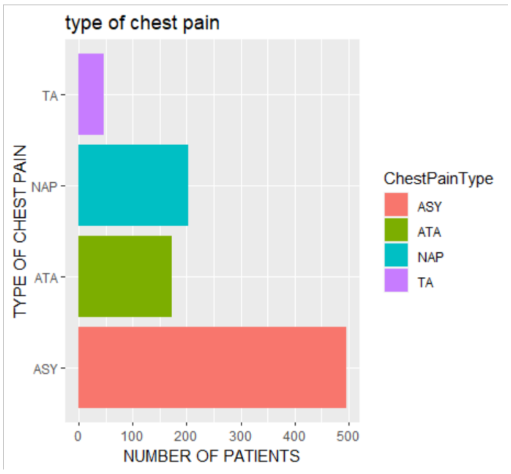

Fig 3: type of pain vs total number of patient

The bar chart shows that many people suffer from asymptomatic pain, but just a few people suffer from traditional angina. Typical angina is caused by physical or mental extortion, whereas asymptomatic pain has no symptoms.

<u>Resting blood pressure:</u> Resting blood pressure is measured in mm Hg.

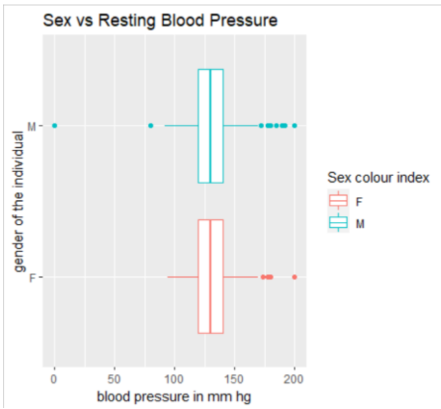

Fig 4: Sex vs Resting Blood Pressure

The blood pressure at resting seems to be the same for both the male and female throughout the observation. However, few outliers can be viewed from the box plot.

<u>Fasting blood sugar:</u> Fasting blood sugar is measured in milligrammes per deciliter. If the blood sugar level is more than 120 mg/dl, the patient is more likely to develop heart disease.

<u>Resting ECG:</u> It is a test that is carried out to detect anomalies in a person. There are numerous forms of resting ECGs, including normal, ST, and LVH (left ventricular hypertrophy).

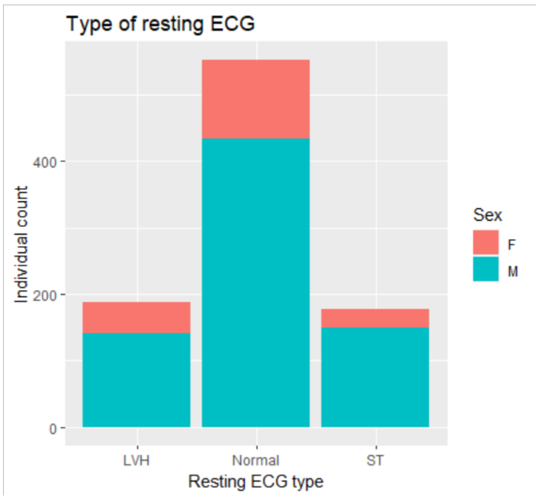

Fig 5: type of Resting ECG

It is evident from the graph that maximum male and female have obtained their ECG result as normal and still been prone to heart disease.

<u>Exercise:</u> Exercise is another parameter that needs to be considered to classify if a person will get cardiac complications. If yes, it means that the person has a good lifestyle, otherwise, the individual is leading a bad lifestyle.

<u>Maximum heart rate:</u> The maximum heart rate achieved by an individual is noted here. The value, however, varies from 60 to 202.

Old peak and ST slope are the other parameters that are given in this dataset. The ST slope is divided into three segments names upsloping, downsloping and flat.

HeartDisease: The heart disease is the column which consists of the output 1 and 0. Here 1 means that the individual has heart disease while 0 means the patient does not have heart disease and he is normal.

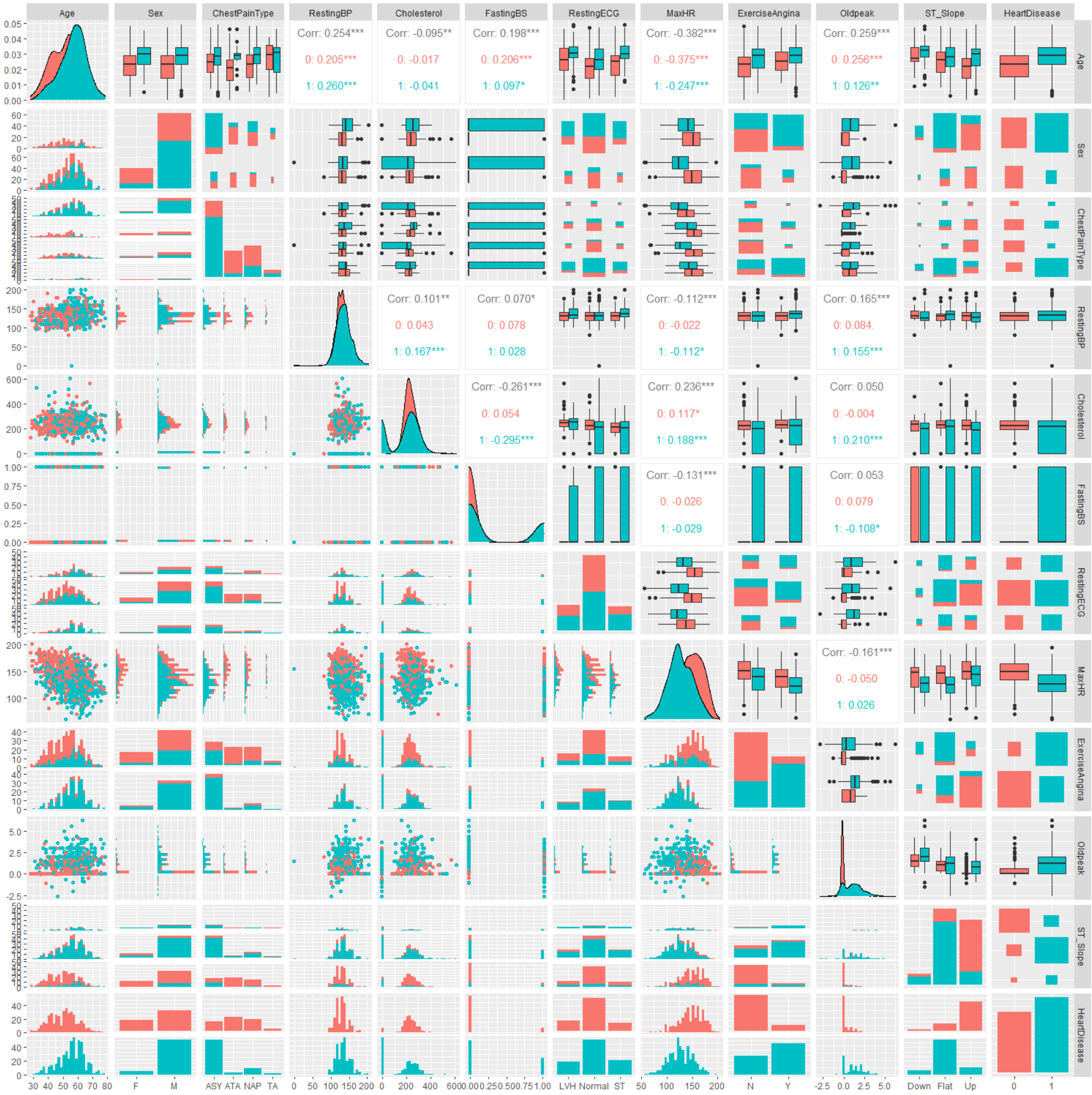

Fig 6: GGPAIR PLOT

The ggpair function in R studios is used to identify the relationship between various variables in the dataset. It gives a basic understanding of the relationship between the variables.

CLASSIFICATION MODEL:

Model 1 – LDA with heart disease as a function of cholesterol, resting blood pressure, and fasting blood sugar.

An LDA function is used to generate the first model. Heart disease is a dependent variable that is influenced by cholesterol, resting blood pressure, and fasting blood sugar levels. A prior probability is nothing more than a null classifier that serves as a benchmark error rate for determining the model's accuracy. The null classifier rate is 56.29 per cent in this case. With the train and test data sets, the LDA model is run, and the relevant data output is obtained.

		Predicted value	
		0	1
Actual value	0	152	98
	1	59	150

Table 1: output of model 1 for the training dataset.

The success rate of the model for the training dataset is 65%

		Predicted value	
		0	1
Actual value	0	133	104
	1	66	156

Table 2: output of model 1 for the test dataset.

With the test data, the model was able to achieve a success rate of around 63 per cent. The model's accuracy has increased by 9%. This is due to the use of three predictors: cholesterol, resting blood pressure and fasting blood sugar. According to this model, 133 people out of 459 will not have heart disease, whereas 156 people will develop heart disease. This model's error rate is estimated to be around 37%. These success rates, however, are not as expected. As a result, in the next model, we utilise different predictors. As the number of predictors utilised grows, so does the accuracy rate.

Null classifier – 56.29%, error percentage – 43.7%

Success rate of the model – 63%, Error rate of the model – 37%

Model 2 - LDA with heart disease as a function of age and maximum heart rate:

Here the LDA function is used to classify the heart diseases based on age and maximum heart rate. The prior probability for the individuals with no heart disease is 45.9% whereas, the prior probability for the individuals with heart disease is 54.03%. The accuracy of the test and train data is then compared using the function.

		Predicted value	
		0	1
Actual value	0	126	68
	1	85	180

Table 3: output of model 2 for the training dataset.

The classification accuracy for the training dataset is 66.6%.

		predicted value	
		0	1
Actual value	0	127	61
	1	72	199

Table 4: output of model 2 for the test dataset.

The classification accuracy for the test dataset is 71%. The classification rate of model 2 has increased by 17%.

Null classifier – 54.03%, error percentage – 45.9%

Success rate of the model – 71%, Error rate of the model – 29%

<u>Model 3: LDA with heart disease as a function of age, fasting blood sugar and maximum heart rate:</u>

Based on age, fasting blood sugar, and maximal heart rate, the LDA function is used to classify heart diseases. The prior probability for the individuals with no heart disease is 46% whereas, the prior probability for the individuals with heart disease is 54%. The function is then called on the test and train data and their accuracy is compared.

		predicted value	
		0	1
Actual value	0	142	70
	1	69	178

Table 5: output of model 3 for the training dataset

The accuracy of the model 3 using the train data set is 69.7%

		predicted value	
		0	1
Actual value	0	134	62
	1	65	198

Table 6: output of model 3 for the test dataset

The classification accuracy for the test dataset is 72.3%. The classification rate of model 3 has increased by 18%.

Null classifier – 54%, error percentage – 46%

Success rate of the model – 72.3%, Error rate of the model – 27.7%

<u>Model 4: LDA with heart disease as a function of age, fasting blood sugar and maximum heart rate and chest pain type:</u>

Based on age, fasting blood sugar, and maximal heart rate, the LDA function is used to classify heart diseases. Furthermore, a categorical variable known as chest pain type is utilised to predict whether a person would develop heart disease. The prior probability for the individuals with no heart disease is 46% whereas, the prior probability for the individuals with heart disease is 54%. The function is then called on the test and train data and their accuracy is compared.

		predicted value	
		0	1
Actual value	0	150	46
	1	61	202

Table 7: output of model 4 for the training dataset

The accuracy of the model 4 using the train data set is 76.6%

		predicted value	
		0	1
Actual value	0	140	32
	1	59	202

Table 8: output of model 4 for the test dataset

The accuracy of model 4 using the test data set is 80.1%.

Null classifier – 54%, error percentage – 46%

Success rate of the model – 80.1%, Error rate of the model – 19.9%

Model 5: LDA with heart disease as a function of age, fasting blood sugar and maximum heart rate and chest pain type, resting ECG:

The LDA function is used to classify the heart diseases based on age, fasting blood sugar, maximum heart rate, chest pain type and resting ECG. The prior probability or null classification model for "0" which is the individuals with no heart disease is 45.9%, and for the individuals with a possibility to get heart disease "1" is 54.03%.

		predicted value	
		0	1
Actual value	0	152	46
	1	59	202

Table 9: output of model 5 for the training dataset

The accuracy of model 5 using the train data set is 77%.

		predicted value	
		0	1
Actual value	0	141	32
	1	58	228

Table 10: output of model 5 for the test dataset

The accuracy of model 5 using the test data set is 80%.

The classification accuracy obtained from model 5 is like the value obtained from model 4. Therefore, resting ECG does not add much value to the classification model.

Null classifier – 54%, error percentage – 46%

Success rate of the model – 80%, Error rate of the model – 20%

Model 6: LDA with heart disease as a function of age, sex, old peak and exercise angina

This is the final model created. Here the heart disease is identified as a function of age, sex, old peak obtained and the exercise angina of an individual. This model is created by neglecting the main predictors like cholesterol and maximum heart rate. The null classifier percentage is 55%.

The dataset is split into test and train data. The code is run separately to identify the classification accuracy of the model.

		predicted value	
		0	1
Actual value	0	176	72
	1	35	176

Table 11: output of model 6 for the training dataset

The accuracy of model 6 using the train data set is 76.6%.

		predicted value	
		0	1
Actual value	0	159	54
	1	40	206

Table 12: output of model 6 for the test dataset

The accuracy of model 6 using the test data set is 79.5%.

The success rate obtained for model 6 is 79.5 %, which is comparatively more than the first 3 models created.

Null classifier – 55%, error percentage – 45%

Success rate of the model – 79.5%, Error rate of the model – 20.5%

Conclusion:

The classification of the dataset is completed in this assignment. Using the linear discriminant analysis technique, the dataset is classified as to whether the patient has heart disease or not. Several LDA models were tested, and the best one was selected. Model 4 has the highest success rate of 80 per cent with the test dataset, hence I believe it is the best of the six models listed. In other words, it says that the algorithm correctly recognises people with heart problems 80% of the time. On the other hand, 20% of the time, it diagnoses the patient inaccurately. This could be problematic in practice because a patient may not be diagnosed with heart disease 20% of the time. This could be changed if more predictors are considered, or if the data is acquired in a more restricted way.

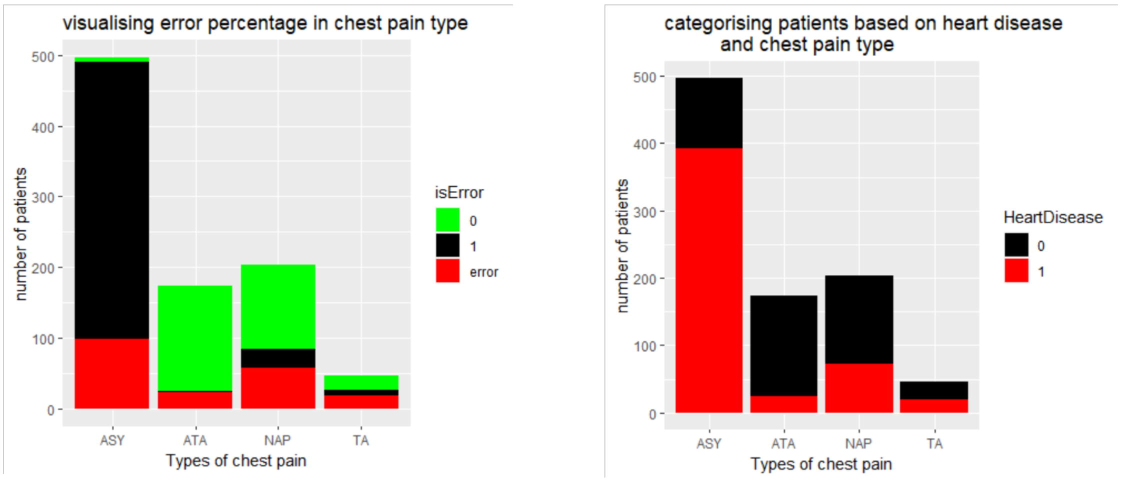

Fig 7: Comparing the errors that occurred in model 4

The graphs above show the inaccuracies that occurred while incorrectly categorising the type of chest pain. It is obvious that approximately 100 people were wrongly diagnosed as having asymptomatic pain, 60 people were falsely labelled as having non-anginal pain, and 20 to 30 people were falsely classified as having atypical angina or typical angina. The error rate is caused by incorrect categorisation. Model 4 has the highest success rate, hence this is plotted for it.

R- CODE:

```
#20 - 05 - 2022 attempt
library(tidyverse)
library(dplyr)
library(MASS)
library(ggplot2)
library(readr)
library(corrplot)

#importing dataset
setwd("C:/CANADA/industrial engineering/summer term/data analytics/dataset")
heartdata= read.csv("heart.csv")
head(heartdata)
str(heartdata)
dim(heartdata)

#cleaning dataset
colSums(is.na(heartdata))
cleandata <- heartdata[complete.cases(heartdata),]
dim(cleandata)
view(cleandata)
str(cleandata)
cleandata$HeartDisease <- as.factor(cleandata$HeartDisease)

library(GGally)
```

```
ggpairs(cleandata, aes(colour = HeartDisease))
attach(cleandata)
str(cleandata)

#linear discriminant analysis
# TRYING MODELS BEFORE SPLITTING THE DATASET INTO TEST AND TRAIN DATA
attach(cleandata)

lda1 = lda(HeartDisease~Cholesterol)
lda1
lda2 = lda(HeartDisease~Cholesterol+RestingBP)
lda2
lda3 = lda(HeartDisease~Cholesterol+RestingBP+FastingBS)
lda3

ldaPred = predict(lda1, cleandata)
ldaPred2 = predict(lda2, cleandata)
ldaPred3 = predict(lda3, cleandata)

#confusion matrix
table(ldaPred$class, cleandata$HeartDisease)
table(ldaPred2$class, cleandata$HeartDisease)
table(ldaPred3$class, cleandata$HeartDisease)

mean(ldaPred$class== cleandata$HeartDisease)
mean(ldaPred2$class== cleandata$HeartDisease)
mean(ldaPred3$class== cleandata$HeartDisease)
view(cleandata)

#spliting training and test data
```

```
IDheart = mutate(cleandata, id=row_number())
View(IDheart)
trainDataSet = sample_frac(IDheart, .5)
View(trainDataSet)
testDataSet = anti_join(IDheart, trainDataSet, by = "id")
View(testDataSet)

#creating model
heartmodel = lda(HeartDisease~Cholesterol+RestingBP+FastingBS , data = trainDataSet)
heartmodel
hearttrainSetPrediction = predict(heartmodel, trainDataSet)
table(hearttrainSetPrediction$class, trainDataSet$HeartDisease)
mean(hearttrainSetPrediction$class== trainDataSet$HeartDisease)

#running model for test data
heartTestSetprediction = predict(heartmodel,testDataSet )
View(heartTestSetprediction$class)
View(testDataSet$HeartDisease)
table(heartTestSetprediction$class , testDataSet$HeartDisease)
mean(heartTestSetprediction$class == testDataSet$HeartDisease)

#creating second model
heartmodel2 = lda(HeartDisease~ Age+MaxHR, data = trainDataSet)
heartmodel2
hearttrainSetPrediction2 = predict(heartmodel2, trainDataSet)
table(hearttrainSetPrediction2$class, trainDataSet$HeartDisease)
mean(hearttrainSetPrediction2$class== trainDataSet$HeartDisease)

#running model for test data
```

```
heartTestSetprediction2 = predict(heartmodel2,testDataSet )
View(heartTestSetprediction2$class)
View(testDataSet$HeartDisease)
table(heartTestSetprediction2$class , testDataSet$HeartDisease)
mean(heartTestSetprediction2$class == testDataSet$HeartDisease)

#creating third model
heartmodel3 = lda(HeartDisease~ Age+MaxHR+FastingBS, data = trainDataSet)
heartmodel3
hearttrainSetPrediction3 = predict(heartmodel3, trainDataSet)
table(hearttrainSetPrediction3$class, trainDataSet$HeartDisease)
mean(hearttrainSetPrediction3$class== trainDataSet$HeartDisease)

#running model for test data
heartTestSetprediction3 = predict(heartmodel3,testDataSet )
View(heartTestSetprediction2$class)
View(testDataSet$HeartDisease)
table(heartTestSetprediction3$class , testDataSet$HeartDisease)
mean(heartTestSetprediction3$class == testDataSet$HeartDisease)

#creating fourth model
heartmodel4 = lda(HeartDisease~ Age+MaxHR+FastingBS+ChestPainType, data = trainDataSet)
heartmodel4
hearttrainSetPrediction4 = predict(heartmodel4, trainDataSet)
table(hearttrainSetPrediction4$class, trainDataSet$HeartDisease)
mean(hearttrainSetPrediction4$class== trainDataSet$HeartDisease)

#running model for test data
heartTestSetprediction4 = predict(heartmodel4,testDataSet )
View(heartTestSetprediction4$class)
View(testDataSet$HeartDisease)
table(heartTestSetprediction4$class , testDataSet$HeartDisease)
mean(heartTestSetprediction4$class == testDataSet$HeartDisease)

#creating fifth model
heartmodel5 = lda(HeartDisease~ Age+MaxHR+FastingBS+ChestPainType+RestingECG,
data = trainDataSet)
heartmodel5
hearttrainSetPrediction5 = predict(heartmodel5, trainDataSet)
table(hearttrainSetPrediction5$class, trainDataSet$HeartDisease)
mean(hearttrainSetPrediction5$class== trainDataSet$HeartDisease)

#running model for test data
heartTestSetprediction5 = predict(heartmodel5,testDataSet )
View(heartTestSetprediction5$class)
View(testDataSet$HeartDisease)
table(heartTestSetprediction5$class , testDataSet$HeartDisease)
mean(heartTestSetprediction5$class == testDataSet$HeartDisease)

#creating sixth model
heartmodel6 = lda(HeartDisease~Age+Sex+Oldpeak+ExerciseAngina, data = trainDataSet)
heartmodel6
hearttrainSetPrediction6 = predict(heartmodel6, trainDataSet)
table(hearttrainSetPrediction6$class, trainDataSet$HeartDisease)
mean(hearttrainSetPrediction6$class== trainDataSet$HeartDisease)

#running model for test data
heartTestSetprediction6 = predict(heartmodel6,testDataSet )
View(heartTestSetprediction6$class)
View(testDataSet$HeartDisease)
table(heartTestSetprediction6$class , testDataSet$HeartDisease)
```

```
mean(heartTestSetprediction6$class == testDataSet$HeartDisease)


#plotting relationships
#histogram for age
ggplot(data= cleandata)+
  geom_histogram(aes(x=Age),
                 position = position_dodge2(padding = .3,
                                            preserve = "single"))+
  ylab("number of patients")+
  xlab("Age group")+
  ggtitle("Age vs heart disease")

#box plot for sex vs cholesterol
ggplot(data = cleandata)+
  geom_boxplot(mapping = aes(y = factor(Sex), x = Cholesterol, colour = Sex))+
  ylab("Sex")+
  xlab("cholesterol in mm/dl")+
  labs(color = "Sex colour index")+
  ggtitle("Sex vs Cholesterol")

#bar plot for chest pain type
ggplot(data=cleandata)+
  geom_bar(mapping = aes(y=ChestPainType, fill = ChestPainType))+
  ylab("TYPE OF CHEST PAIN")+
  xlab("NUMBER OF PATIENTS")+
  ggtitle("type of chest pain")


#box plot for sex vs Resting Blood Pressure
ggplot(data = cleandata)+
  geom_boxplot(mapping = aes(y = factor(Sex), x = RestingBP, colour = Sex))+
  ylab("gender of the individual")+
  xlab("blood pressure in mm hg")+
  labs(color = "Sex colour index")+
  ggtitle("Sex vs Resting Blood Pressure")

#resting graph
ggplot(data=cleandata)+
  geom_bar(mapping = aes(x= RestingECG, fill = Sex))+
  ylab("Individual count")+
  xlab("Resting ECG type")+
  ggtitle("Type of resting ECG")

#creating conclusion graph to visualize the error.
attach(cleandata)
model1D = lda(HeartDisease~ Age+MaxHR+FastingBS+ChestPainType)
guess1D = predict(model1D, cleandata)
heartwith1Dguesses = mutate(cleandata, guess=guess1D$class,
                            isError = ifelse(guess==HeartDisease, as.character(guess),
                                             "error"))
View(heartwith1Dguesses)

ggplot(data= heartwith1Dguesses )+geom_bar( aes(x= ChestPainType,
fill=isError))+scale_fill_manual(values = c("green", "black", "red", "blue"))+
  ylab("number of patients")+
  xlab("Types of chest pain")+
  labs(color = "outcomes")+
  ggtitle("visualising error percentage in chest pain type")

ggplot(data=cleandata)+geom_bar(aes(x= ChestPainType,
                                fill=HeartDisease))+scale_fill_manual(values = c( "black", "red",
"blue"))+
  ylab("number of patients")+
  xlab("Types of chest pain")+
  labs(color = " heart disease outcomes")+
  ggtitle("categorising patients based on heart disease
          and chest pain type")
```

REFERENCE:

1. fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved [Date Retrieved] from https://www.kaggle.com/fedesoriano/heart-failure-prediction.